First Hit Fwd Refs Previous Doc Next Doc Go to Doc#

End of Result Set

☑ Generate Collection Print

L4: Entry 5 of 5 File: USPT Mar 14, 2000

DOCUMENT-IDENTIFIER: US 6038527 A
TITLE: Method for generating descriptors for the classification of texts

Abstract Text (1):
The proposed method for generating descriptors for the classification of texts provides a breakdown of more complex word forms by way of matching with the entirety of word forms occurring within a compilation of training texts. No morphological or linguistic knowledge base is necessary for the preferably cyclically continued breakdown, nor for the accompanying drawing up of stop word prefix and suffix lists. Simple morphological knowledge is provided by prescribing minimum requirements with respect to the form of descriptors and text sections. The method is particularly flexible and can be easily adapted to new applications. The method is also very error-tolerant and thus particularly suited for the classification of digitized texts which are produced from written texts by means of character recognition methods or from spoken texts by means of language recognition methods.

Brief Summary Text (3):
The classification of a text is an assignment to a specific text class and is an important preprocessing step for the automatic further processing of texts. In particular for the automatic interpretation of texts, a preceding classification is of considerable importance because in this manner the expenditure for the knowledge base which needs to be maintained such as, e.g., dictionary memory, syntactic and semantic structure definition, can be limited considerably and the recognition performance can be greatly increased.

Brief Summary Text (4):
Text classification can be divided roughly into two steps, namely the extraction of descriptors and, based on this, the assignment to a class. The selection of the descriptors is of essential importance. The selection is a problem especially for natural language texts having a variety of word forms.

Brief Summary Text (5):
For texts in the English language, which has a small morphological variation, the use of complete word forms or phrases is proposed in "Feature Selection and Feature Extraction for Text Categorization" by D. Lewis in Proc. of Speech and Natural Language Workshop 1992. For classification tasks in morphologically richer languages, word segments can be used as descriptors, with, e.g., the text being broken down into n-grams in "N-Gram-Based Text Categorization" by Canvar/Trenkle in Proc. of Int. Symp. on Document Analysis and Information Retrieval 1994, or use of a reduction to basic forms in "Using IR Techniques for Text Classification in Document Analysis" by R. Hoch in Proc. of SIGIR, 1994.

Detailed Description Text (9):
The beginning and ending character sequences comprised in the prefix and suffix lists are separated from the word forms and word segments present after the breakdown. In this manner, mainly different variation forms of root words can be reduced to their word stem and be combined therein. For the separation of prefixes and suffixes, a frame is also set advantageously for admissible separations by predetermining minimum requirements with respect to the word segments remaining after the separation, e.g., the descriptor restrictions mentioned for the initial exclusion of unsuitable character sequences. The character sequences (word forms and word segments) that are left after breakdown and separation are considered to be suitable descriptors and are used further. The separation of prefixes and suffixes does not have to be limited to the method phase after the completion of all breakdowns, rather it can also be carried out alternatively or additionally in intermediate stages. The reduction of the word forms by way of breakdown or separation does not require any morphological knowledge or

with the specification of minimum requirements only very simple morphological knowledge. This means that in the grammatical sense, faulty analyses and separations are admissible and, as a rule, do occur. Therefore, a word stem is often no longer recognizable in the word forms or word segments that are left. They can also again assume the form of initially excluded stop words. These word truncations are also equally suited as descriptors since they are specific for the text types forming the training texts by virtue of their derivation from the training texts and since they function in the same manner for the training of the classifier as well as for the classification of unknown texts.

Detailed Description Text (12):
The list of descriptors whose use should be continued and the stop word list form the basis for the subsequent text classification in the training phase of the classifier as well as in the classification of unknown texts. Classifiers are generally known from the prior art.

Detailed Description Text (64):
While, in the first example, which was construed for the sake of clarity, the descriptors still bear some resemblance to the underlying word stems, the descriptors often seem to not bear any relation to content in the actual case. This clarifies the difference between the substantially statistical procedure which is advantageous for automatic processing compared to a text analysis on a linguistic knowledge base. The ability to also classify texts containing many errors, as in the example specified above, illustrates the high error tolerance of the method according to the invention.

Other Reference Publication (2):
R. Hoch: "Using IR techniques for text classification in document analysis". In: SIGIR '94. Jul. 3-6, 1994, Dublin, Ireland, pp. 31-40.

Previous Doc Next Doc Go to Doc#

First Hit Fwd Refs Previous Doc Next Doc Go to Doc#

☑ Generate Collection Print

L4: Entry 3 of 5 File: USPT May 27, 2003

DOCUMENT-IDENTIFIER: US 6571225 B1
TITLE: Text categorizers based on regularizing adaptations of the problem of computing linear separators

Brief Summary Text (15):
It is also known that the generalization performance of a linear classifier trained to minimize the training error is determined by its capacity, which can be measured by the concept of covering number, originally studied by A. N. Kolmogorov and V. M. Tihomirov, ".epsilon.-entropy and .epsilon.-capacity of sets in functional spaces", Amer. Math. Soc. Transl., 17(2):277-364 (1961). In learning theory, the VC-dimension is used to bound the growth rate of covering numbers as a function of sample size. It can be shown that the average generalization performance of a linear classifier obtained from minimizing training error is O(d/n)more than the optimal generalization error when the training set consisted of n examples. (The notation O here indicates that the hidden factor may have a polynomial dependence on log(n).) Clearly, if d is large as compared to n, then the generalization performance from the perceptron algorithm will be poor. Unfortunately, large dimensionality is typical for many real-word problems such as text-classification problems, which can have tens of thousands of features. Vapnik realized that by using regularization techniques originated from the numerical solution of ill-posed systems, as described, for example, by A. N. Tikhonov and V. Y. Arsenin in Solution of Ill-Posed Problems, W. H. Winston, Washington, D.C. (1977), one can avoid the dimensional dependency and thus achieve better generalization performance for certain problems, as described in V. N. Vapnik, Estimation of Dependencies Based on Empirical Data, Springer-Verlag, New York (1982), translated from the Russian by Samuel Kotz, and V. N. Vapnik, The Nature of Statistic Learning Theory, Springer-Verlag, New York (1995).

Detailed Description Text (12):
One part of this method is use of a computer to carry out document tokenization, as shown in function block 21. "Tokenization" means extracting a sequence of words or tokens from a sequence of characters. This is shown in detail in FIG. 3. This functionality is common to most methods of text categorization. First, a document d is read in function block 30. The document d is segmented into sections, if any, whose separate identity is significant for categorization in function block 31. For instance, it may be desirable to keep the header separate from the body of a document, in which case, a word extracted from the header would give rise to a different feature from the same word extracted from the body. Each section that contains text is tokenized in function block 32. The steps shown in function blocks 33 and 34 are optional, however, executing these steps may improve performance. All tokens are converted to canonical forms, i.e., stemming, in function block 33. "Stemming" is the replacement of a word by its stem or canonical form, as in replacing plural nouns with singular nouns. Stop words, i.e., common words not useful for categorization, are deleted in function block 34. For example, articles "a", "an", and "the" and prepositions are often found in stop word lists. If both steps (blocks 33 and 34) are performed, they may be performed in either order, because neither has logical precedence over the other, as long as one realizes that the right stop word list to use may be affected by whether stop word elimination comes before or after stemming. Also, the elimination of stop words may in some instances be logically subsumed by subsequent feature selection, to be discussed below.

Previous Doc Next Doc Go to Doc#